Coarse-to-Fine Latent Diffusion for Pose-Guided Person Image Synthesis

Yanzuo Lu[1] Manlin Zhang[1] Andy J Ma[1,2,3 *] Xiaohua Xie[1,2,3] Jianhuang Lai[1,2,3,4]

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]Guangdong Province Key Laboratory of Information Security Technology, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
[4]Pazhou Lab (HuangPu), Guangzhou, China

{luyz5, zhangmlin3}@mail2.sysu.edu.cn, {majh8, xiexiaoh6, stsljh}@mail.sysu.edu.cn

Abstract

Diffusion model is a promising approach to image generation and has been employed for Pose-Guided Person Image Synthesis (PGPIS) with competitive performance. While existing methods simply align the person appearance to the target pose, they are prone to overfitting due to the lack of a high-level semantic understanding on the source person image. In this paper, we propose a novel Coarse-to-Fine Latent Diffusion (CFLD) method for PGPIS. In the absence of image-caption pairs and textual prompts, we develop a novel training paradigm purely based on images to control the generation process of the pre-trained text-to-image diffusion model. A perception-refined decoder is designed to progressively refine a set of learnable queries and extract semantic understanding of person images as a coarse-grained prompt. This allows for the decoupling of fine-grained appearance and pose information controls at different stages, and thus circumventing the potential overfitting problem. To generate more realistic texture details, a hybrid-granularity attention module is proposed to encode multi-scale fine-grained appearance features as bias terms to augment the coarse-grained prompt. Both quantitative and qualitative experimental results on the DeepFashion benchmark demonstrate the superiority of our method over the state of the arts for PGPIS. Code is available at https://github.com/YanzuoLu/CFLD.

1. Introduction

Pose-Guided Person Image Synthesis (PGPIS) aims to translate the source person image into a specific target pose while preserving the appearance as much as possible. It has a wide range of applications, including film production, virtual reality, and fashion e-commerce. Most existing methods along this line are developed based on Generative Adversarial Networks (GANs) [6, 17, 19, 22–25, 30, 31, 35, 36, 39, 46, 48, 53, 54]. Nevertheless, the GAN-based approach may suffer from the instability of min-max training objective and difficulty in generating high-quality images in a single forward pass.

As a promising alternative to GANs for image generation, diffusion models synthesize more realistic images progressively from a series of denoising steps. The recently prevailing text-to-image latent diffusion model, such as Stable Diffusion (SD) [32] may now generate compelling person images conditioned on a given textual prompt. The appearance of the generated person can be determined by well-designed prompts [18, 28] or prompt learning [50, 51].

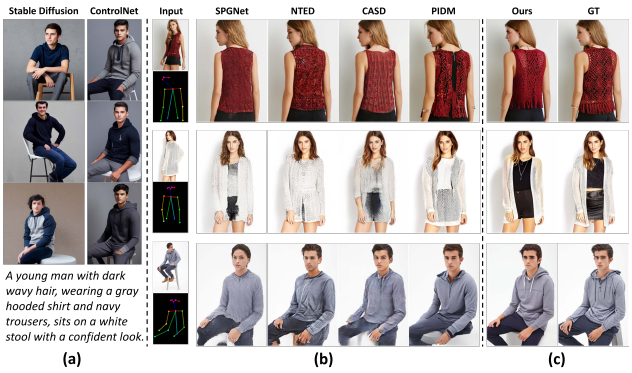

Figure 1. (a) The appearance of person image varies significantly given only a textual prompt for image generation by using Stable Diffusion [32] or ControlNet [47] with OpenPose guidance [3]. (b) Simply aligning the source appearance to the target pose without a semantic understanding of person image can easily lead to overfitting, such that the generated images become distorted and unnatural. (c) Our method learns the coarse-grained prompt for a comprehensive perception of the source image and injects fine-grained appearance features as bias terms, thus generating high-quality images with better generalization performance.

*Corresponding author.

With more reliable structural guidance [26, 47], the synthesized person images can be further constrained to specific poses. Though the text-to-image diffusion generates realistic images from textual prompts with high-level semantics, its training paradigm requires extensive image-caption pairs that are labor-expensive to collect for PGPIS. More importantly, due to the differing information densities between language and vision [11], even the most detailed textual descriptions inevitably introduce ambiguity and may not accurately preserve the appearance as illustrated in Fig. 1(a).

More recently, several diffusion-based approaches have emerged for PGPIS. A texture diffusion module is proposed by PIDM [1] to model the complex correspondence between the appearance of source image and the target pose. Since the denoising process at the high-resolution pixel level is computationally expensive, PoCoLD [9] reduces both the training and inference costs by mapping pixels to low-dimensional latent spaces with a pre-trained Variational Autoencoder (VAE) [7]. In PoCoLD, the correspondence is further exploited by a pose-constrained attention module based on additional 3D Densepose [8] annotations. While both the PIDM and PoCoLD generate more realistic texture details by aligning the source image to the target pose, they lack **a high-level semantic understanding of person images**. Therefore, they are prone to overfitting and poor generalization performance when synthesizing exaggerated poses that are vastly different from the source image or rare in the training set. As demonstrated in Fig. 1(b), the generated images become distorted and unnatural in these cases, which is in line with several GAN-based approaches.

In this work, we propose a novel Coarse-to-Fine Latent Diffusion (CFLD) method for PGPIS. Our approach breaks the conventional training paradigm which leverages textual prompts to control the generation process of a pre-trained SD model. Instead of conditioning on the human-generated signals, i.e. languages that are highly semantic and information-dense, we facilitate a coarse-to-fine appearance control method purely based on images. To obtain the aforementioned semantic understanding specific to person images, we endeavor to decouple the fine-grained appearance and pose information controls at different stages by introducing a *perception-refined decoder*. The perception of the source person image is achieved by randomly initializing a set of learnable queries and progressively refining them in the following decoder blocks via cross-attention. The decoder output serves as a coarse-grained prompt to describe the source image, focusing on the common semantics across different person images, e.g. human body parts and attributes such as age and gender. Moreover, we design a *hybrid-granularity attention* module to effectively encode multi-scale fine-grained appearance features as bias terms to augment the coarse-grained prompt. In this way, the source image is able to be aligned with the target pose by supplementing only the necessary fine-grained details under the guidance of the coarse-grained prompt, thus achieving better generalization as illustrated in Fig. 1(c).

Our main contributions can be summarized as follows,

- We present a novel training paradigm in the absence of image-caption pairs to overcome the limitations when applying text-to-image diffusion to PGPIS. We propose a perception-refined decoder to extract semantic understanding of person images as a coarse-grained prompt.
- We formulate a new hybrid-granularity attention module to bias the coarse-grained prompt with fine-grained appearance features. Thus, the texture details of generated images are better controlled and become more realistic.
- We conduct extensive experiments on the DeepFashion [20] benchmark and achieve the state-of-the-art performance both quantitatively and qualitatively. User studies and ablations validate the effectiveness of our method.

2. Related Work

2.1. Pose-Guided Person Image Synthesis

Ma *et al*. [23] first presents the task of pose-guided person image synthesis and refines the generated images in an adversarial manner. To decouple the pose and appearance information, early approaches [6, 24] propose to learn pose-irrelevant features but fail to handle the complex texture details with vanilla convolutional neural networks. To alleviate this problem, auxiliary information is introduced to improve the generation quality, such as parsing [25] and UV maps [35]. Recent approaches [17, 19, 30, 31, 39, 54] focus on modeling the spatial correspondence between pose and appearance, with the more frequent use of parsing maps [22, 46, 53]. PIDM [1] and PoCoLD [9] are developed based on diffusion models to prevent from the drawbacks in the generative adversarial networks, including the instability of min-max training objective and difficulty in synthesising high-resolution images. Both of these two diffusion-based methods extend the idea of spatial correspondence to model the relation between the appearance of source image and target pose via the cross-attention mechanism. We argue this leads to overfitting by simply aligning the source appearance to the target pose without a high-level semantic understanding of the person image. In this work, we propose a coarse-to-fine latent diffusion method by incorporating both the coarse-grained prompt and fine-grained appearance features. The coarse-grained prompt is gradually refined to give a semantic perception, while the fine-grained appearance features are injected as bias terms to control texture details in our proposed hybrid-granularity attention module. Both the quantitative and qualitative experiments validate that our method achieves state-of-the-art results with stronger generalization performance.

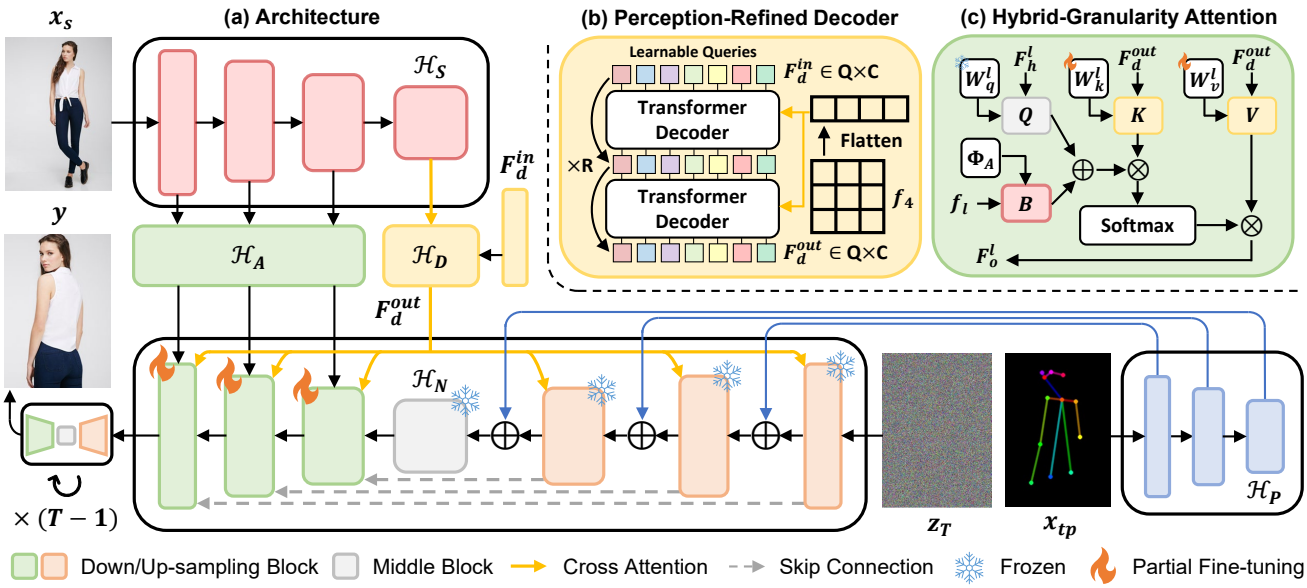

Figure 2. (a) Architecture of our proposed Coarse-to-Fine Latent Diffusion (CFLD) method. For pose-guided latent diffusion, we incorporate a lightweight pose adapter $\mathcal{H}_P$ from [26] to add its output feature maps to the end of each down-sampling block of the pre-trained UNet $\mathcal{H}_N$ for efficient structural guidance. To achieve a coarse-to-fine appearance control, we propose a perception-refined decoder $\mathcal{H}_D$ and hybrid-granularity attention module $\mathcal{H}_A$, both of which take the multi-scale feature maps from a source image encoder $\mathcal{H}_S$ as inputs. (b) The coarse-grained prompt is obtained by refining the learnable queries progressively in our proposed $\mathcal{H}_D$. (c) We encode the multi-scale fine-grained appearance features as bias terms in the up-sampling blocks for better texture details within $\mathcal{H}_A$.

2.2. Controllable Diffusion Models

Diffusion models have recently emerged and demonstrated their potential for high-resolution image synthesis. The core idea is to start with a simple noise vector and gradually transform it into a high-quality image through multiple denoising iterations. Beyond unconditional generation [15, 37, 38], various methods have been introduced to incorporate user-supplied control signals into the generation process, enabling more controllable image generation. For instance, [5] introduces the usage of classifier gradients to condition on the generation, while [14] proposes a classifier-free control mechanism employing a weighted summation of conditional and unconditional outputs for controllable synthesis. Moreover, the Latent Diffusion Model (LDM) performs diffusion in the latent space and injects the conditioning signals via a specific encoder and cross-attention. Building upon the pre-trained LDM like Stable Diffusion (SD) [32], subsequent works have explored to bias the latent space by adding extra controls [26, 47], as well as further to provide users with control over the generated content [12, 40]. Rather than employing a high-level conditioning prompt throughout the generation, we design a coarse-to-fine conditioning process that adjusts the latent features at different stages within the UNet-based prediction network, providing better controllable pose-guided person image synthesis.

3. Method

3.1. Preliminary

Our method builds on top of the text-to-image latent diffusion model, i.e., Stable Diffusion (SD) [32] with high-quality image generation ability. There are two main stages in the SD model: a Variational Autoencoder (VAE) [7] that maps between raw-pixel space and low-dimensional latent space and an UNet-based prediction model [33] for denoising diffusion image generation. It follows the general idea of Denoising Diffusion Probabilistic Model (DDPM) [15], which formulates a forward diffusion process and a backward denoising process of $T = 1000$ steps. The diffusion process progressively adds random Gaussian noise $\epsilon \sim \mathcal{N}(0, \boldsymbol{I})$ to the initial latent $\boldsymbol{z}_0$, mapping it into noisy latents $\boldsymbol{z}_t$ at different timesteps $t \in [1, T]$,

$$\boldsymbol{z}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \tag{1}$$

where $\bar{\alpha}_1, \bar{\alpha}_2, ..., \bar{\alpha}_T$ are derived from a fixed variance schedule. The denoising process learns the UNet $\epsilon_\theta(\boldsymbol{z}_t, t, \boldsymbol{c})$ to predict the noise and reverse this mapping, where $\boldsymbol{c}$ is the conditional embedding output by e.g. the CLIP [29] text encoder in [32]. The optimization can be formulated as,

$$\mathcal{L}_{mse} = \mathbb{E}_{\boldsymbol{z}_0,\boldsymbol{c},\epsilon,t}\left[\|\epsilon - \epsilon_\theta(\boldsymbol{z}_t, t, \boldsymbol{c})\|_2^2\right]. \tag{2}$$

3.2. Coarse-to-Fine Latent Diffusion

Architecture and Overview. Fig. 2(a) shows the architecture of our proposed method. For concise illustration, we omit the encoder $\mathcal{E}$ and decoder $\mathcal{D}$ of the VAE [7] model in this figure. In the training phase, we are given sets of the source image $\boldsymbol{x}_s$, source pose $\boldsymbol{x}_{sp}$, target pose $\boldsymbol{x}_{tp}$, and ground-truth image $\boldsymbol{x}_g$. The source image passes through an image encoder $\mathcal{H}_S$ (e.g. swin transformer [21]), from which we extract a stack of multi-scale feature maps $\boldsymbol{F}_s = [\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3, \boldsymbol{f}_4]$ for a coarse-to-fine *appearance control*. The coarse-grained prompts are learned by our Perception-Refined Decoder (PRD) $\mathcal{H}_D$ and serve as conditional embeddings in both down-sampling and up-sampling blocks of the UNet $\mathcal{H}_N$. While the down-sampling block in $\mathcal{H}_N$ remains intact in our method, we reformulate the up-sampling block with our Hybrid-Granularity Attention module (HGA) $\mathcal{H}_A$ to bias the coarse-grained prompt with fine-grained appearance features for more realistic textures. More details about $\mathcal{H}_D$ and $\mathcal{H}_A$ will be presented later.

For efficient *pose control*, we adopt a lightweight pose adapter $\mathcal{H}_P$ that consists of several ResNet blocks [10]. The output feature maps of $\mathcal{H}_P$ are added directly to the end of each down-sampling block as in [26]. This requires no additional fine-tuning and explicitly decouples the fine-grained appearance and pose information controls. At different scales of down-sampling, the pose information is only aligned with the same coarse-grained prompts given by our PRD as conditional embeddings, rather than the different multi-scale fine-grained appearance features in the common practice [1, 9]. In this way, the HGA module learns all the pose-irrelevant texture details at the up-sampling stage and is not prone to overfitting. Denote the initial latent state for the ground-truth image as $\boldsymbol{z}_0 = \mathcal{E}(\boldsymbol{x}_g)$. The MSE loss in Eq. (2) is thus rewritten as,

$$\mathcal{L}_{mse} = \mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{x}_s, \boldsymbol{x}_{tp}, \epsilon, t}\left[\|\epsilon - \epsilon_\theta(\boldsymbol{z}_t, t, \boldsymbol{x}_s, \boldsymbol{x}_{tp})\|_2^2\right]. \quad (3)$$

Perception-Refined Decoder. Instead of utilizing multi-scale appearance features as conditional embeddings as in the existing diffusion-based approaches [1, 9], we propose to decouple the controls from the fine-grained appearance and pose information at different stages. Thus we design a Perception-Refined Decoder (PRD) to extract semantic understanding of person images as a coarse-grained prompt, given the flattened last-scale output $\boldsymbol{f}_4$ from $\mathcal{H}_S$ as illustrated in Fig. 2(b). By revisiting how people perceive a person image, we find several common characteristics, i.e., human body parts, age, gender, hairstyle, clothing, and so on, as demonstrated in Fig. 1(a). This inspires us to maintain a set of learnable queries $\boldsymbol{F}_d^{in} \in \mathbb{R}^{Q \times D}$ representing different semantics of person images. They are randomly initialized and progressively refined with the standard transformer decoders [42]. The source image conditioning $\boldsymbol{f}_4$ interacts via the cross-attention module at each decoder block. After R blocks of refinement, we obtain the coarse-grained prompt $\boldsymbol{F}_d^{out}$, which serves as the conditional embedding and inputs to both down-sampling and up-sampling in $\mathcal{H}_N$.

Hybrid-Granularity Attention. To precisely control the texture details of generated images, we introduce the Hybrid-Granularity Attention module (HGA) that is embedded in different scales ($l \in \{1, 2, 3\}$) of up-sampling blocks in $\mathcal{H}_N$, where we refer $\boldsymbol{F}_h^l$, $\boldsymbol{F}_o^l$ to its input and output. Given the multi-scale feature maps $\boldsymbol{f}_l$ of the source image from $\mathcal{H}_S$, the HGA module aims to compensate for the missing necessary details in the coarse-grained prompts. To achieve this, we formulate the HGA module that naturally follows a coarse-to-fine learning curriculum.

Specifically, we propose to inject multi-scale texture details by biasing the queries of cross-attention in the up-sampling blocks as shown in Fig. 2(c), i.e.,

$$\begin{aligned} &\boldsymbol{Q} = \boldsymbol{W}_q^l \boldsymbol{F}_h^l, \quad \boldsymbol{K} = \boldsymbol{W}_k^l \boldsymbol{F}_d^{out}, \quad \boldsymbol{V} = \boldsymbol{W}_v^l \boldsymbol{F}_d^{out}, \\ &\boldsymbol{B} = \phi_A(\boldsymbol{f}_l), \quad \boldsymbol{F}_o^l = \text{softmax}(\frac{(\boldsymbol{Q} + \boldsymbol{B})\boldsymbol{K}^T}{\sqrt{d}})\boldsymbol{V}, \end{aligned} \quad (4)$$

where $\boldsymbol{W}_q^l, \boldsymbol{W}_k^l, \boldsymbol{W}_v^l$ are specific projection layers for the l-th scale up-sampling block of dimension d. ϕ_A is a fine-grained appearance encoder that mainly consists of K transformer layers with a zero convolution [47] added in the beginning and the end. The zero convolution is a standard 1×1 convolution layer with both weight and bias initialized as zeros. It keeps the gradient of $\mathcal{H}_A$ back to $\mathcal{H}_S$ small enough in the early stage of training, so that the image encoder $\mathcal{H}_S$ and the more easily converged perception-refined decoder $\mathcal{H}_D$ can focus on learning to provide a high-level semantic understanding compatible with the pre-trained SD model. Since we have decoupled the controls of the fine-grained appearance and pose information at different stages, the target pose can be well controlled without overfitting during the down-sampling process. Therefore, such a design encourages the HGA module to slowly fill in more fine-grained textures to better align the generation with the source image during training. Note that $\boldsymbol{W}_k^l, \boldsymbol{W}_v^l$ in the up-sampling blocks are trainable parameters. They are the only trainable parameters of the entire $\mathcal{H}_N$, which accounts for only **1.2%** of all the parameters in the pre-trained SD model.

3.3. Optimization

To assist the source-to-target pose translation, we follow the insights in [48] to conduct source-to-source self-reconstruction for training. The reconstruction loss is,

$$\mathcal{L}_{rec} = \mathbb{E}_{\boldsymbol{z}_0, \boldsymbol{x}_s, \boldsymbol{x}_{sp}, \epsilon, t}\left[\|\epsilon - \epsilon_\theta(\boldsymbol{z}_t, t, \boldsymbol{x}_s, \boldsymbol{x}_{sp})\|_2^2\right], \quad (5)$$

where $\boldsymbol{z}_0 = \mathcal{E}(\boldsymbol{x}_s)$ and $\boldsymbol{z}_t$ is the noisy latent mapped from $\boldsymbol{z}_0$ at timestep t. The overall objective is written as,

$$\mathcal{L}_{overall} = \mathcal{L}_{mse} + \mathcal{L}_{rec}. \quad (6)$$

Component	Default	Trainable Params.
$\mathcal{H}_S$	Swin-B [21]	87.0M
$\mathcal{H}_A$	$K = 4$	22.5M
$\mathcal{H}_D$	$R = 8, Q = 16, C = 768$	97.7M
$\mathcal{H}_P$	Adapter [26]	30.6M
$\mathcal{H}_N$	up-sampling $\boldsymbol{W}_k^l, \boldsymbol{W}_v^l$	10.3M

Method	Pose Info. & Annotation	Training Epochs	Trainable Params.
PIDM [1]	2D OpenPose [3]	300	688.0M
PoCoLD [9]	3D DensePose [8]	100	395.9M
CFLD (Ours)	2D OpenPose [3]	100	**248.2M**

Table 1. The default settings and the number of trainable parameters in each component of our method and comparison with other diffusion-based methods.

Moreover, we adopt the cubic function $t = (1 - (\frac{t}{T})^3) \times T,\ t \in \text{Uniform}(1, T)$ for the distribution of timestep t. It increases the probability of t falling in the early sampling stage and strengthens the guidance, which helps to converge faster and shorten the training time.

Sampling. Once the conditional latent diffusion model is learned, the inference can be performed and starts by sampling a random Gaussian noise $\boldsymbol{z}_T \sim \mathcal{N}(0, \boldsymbol{I})$. The predicted latent $\tilde{\boldsymbol{z}}_0$ is obtained by reversing the schedule in Eq. (1) using the denoising network ϵ_t at each timestep $t \in [1, T]$. We adopt the cumulative classifier-free guidance [2, 9, 14] to strengthen both the source appearance and target pose guidance, i.e.,

$$\begin{aligned}\epsilon_t = &\epsilon_\theta(\boldsymbol{z}_t, t, \varnothing, \varnothing) \\ &+ w_{\text{pose}}(\epsilon_\theta(\boldsymbol{z}_t, t, \varnothing, \boldsymbol{x}_{tp}) - \epsilon_\theta(\boldsymbol{z}_t, t, \varnothing, \varnothing)) \\ &+ w_{\text{app}}(\epsilon_\theta(\boldsymbol{z}_t, t, \boldsymbol{x}_s, \boldsymbol{x}_{tp}) - \epsilon_\theta(\boldsymbol{z}_t, t, \varnothing, \boldsymbol{x}_{tp})).\end{aligned} \tag{7}$$

When the source image $\boldsymbol{x}_s$ is missing, we use learnable vectors as the conditional embeddings. The learnable vectors are trained with a probability of $\eta\%$ to drop both $\boldsymbol{x}_s$ and $\boldsymbol{x}_p$ during training. The outputs of the pose adapter $\mathcal{H}_P$ will be set to all zeros if the target pose $\boldsymbol{x}_{tp}$ is missing. We use the DDIM scheduler [37] to speed up the sampling with 50 steps as the same as in [1, 9]. Finally, the generated image is obtained by the VAE decoder $\boldsymbol{y} = \mathcal{D}(\tilde{\boldsymbol{z}}_0)$.

4. Experiments

4.1. Setup

Dataset. We follow [9, 31] to conduct experiments on the In-Shop Clothes Retrieval benchmark of DeepFashion [20] and evaluate on both the 256×176 and 512×352 resolutions. This dataset consists of 52,712 high-resolution person images in the fashion domain. The dataset split is the same as in PATN [54], where 101,966 and 8,570 non-overlapping pairs are selected for training and testing, respectively.

Method	Venue	FID↓	LPIPS↓	SSIM↑	PSNR↑
Evaluate on 256×176 resolution					
PATN [54]	CVPR 19'	20.728	0.2533	0.6714	-
ADGAN [25]	CVPR 20'	14.540	0.2255	0.6735	-
GFLA [30]	CVPR 20'	9.827	0.1878	0.7082	-
PISE [46]	CVPR 21'	11.518	0.2244	0.6537	-
SPGNet† [22]	CVPR 21'	16.184	0.2256	0.6965	17.222
DPTN† [48]	CVPR 22'	17.419	0.2093	0.6975	17.811
NTED† [31]	CVPR 22'	8.517	0.1770	0.7156	17.740
CASD† [53]	ECCV 22'	13.137	0.1781	0.7224	17.880
PIDM† [1]	CVPR 23'	6.812	0.2006	0.6621	15.630
PIDM‡ [1]	CVPR 23'	6.440	0.1686	0.7109	17.399
PoCoLD* [9]	ICCV 23'	8.067	0.1642	0.7310	-
CFLD (Ours)		**6.804**	**0.1519**	**0.7378**	**18.235**
VAE Reconstructed		7.967	0.0104	0.9660	33.515
Ground Truth		7.847	0.0000	1.0000	$+\infty$
Evaluate on 512×352 resolution					
CoCosNet2 [52]	CVPR 21'	13.325	0.2265	0.7236	-
NTED† [31]	CVPR 22'	7.645	0.1999	0.7359	17.385
PoCoLD* [9]	ICCV 23'	8.416	0.1920	0.7430	-
CFLD (Ours)		**7.149**	**0.1819**	**0.7478**	**17.645**
VAE Reconstructed		8.187	0.0217	0.9231	30.214
Ground Truth		8.010	0.0000	1.0000	$+\infty$

Table 2. Quantitative comparisons with the state of the arts in terms of image quality. † We strictly follow the evaluation implementation in NTED [31] and reproduce these results based on the checkpoints rather than generated images provided by the authors for a fair comparison. ‡ Results are obtained using the generated images released by the authors. * Results are cited from PoCoLD [9] without publicly available checkpoints and generated images. Others are cited from NTED.

Objective metrics. We use four different metrics to evaluate the generated images quantitatively, including FID [13], LPIPS [49], SSIM [44] and PSNR. Both FID and LPIPS are based on deep features. The Fréchet Inception Distance (FID) calculates the Wasserstein-2 distance [41] between the distributions of generated and real images using Inception-v3 [34] features, and the Learned Perceptual Image Patch Similarity (LPIPS) leverages a network trained on human judgments to measure reconstruction accuracy in the perceptual domain. As for the Structural Similarity Index Measure (SSIM) and Peak Signal to Noise Ratio (PSNR), they quantify the similarity between generated images and ground truths at the pixel level.

Subjective metrics. In addition to the objective metrics, we follow [9] to use the Jab [1, 22, 53] metric in our user study to calculate the percentage of generated images that were considered the best among all methods [1, 22, 31, 48, 53]. Moreover, in order to measure the similarity between the generated images and real data, we quantify the R2G and G2R metrics as many early approaches did [23, 36, 54]. R2G represents the percentage of real images considered as generated and G2R represents the percentage of generated images considered as real by humans.

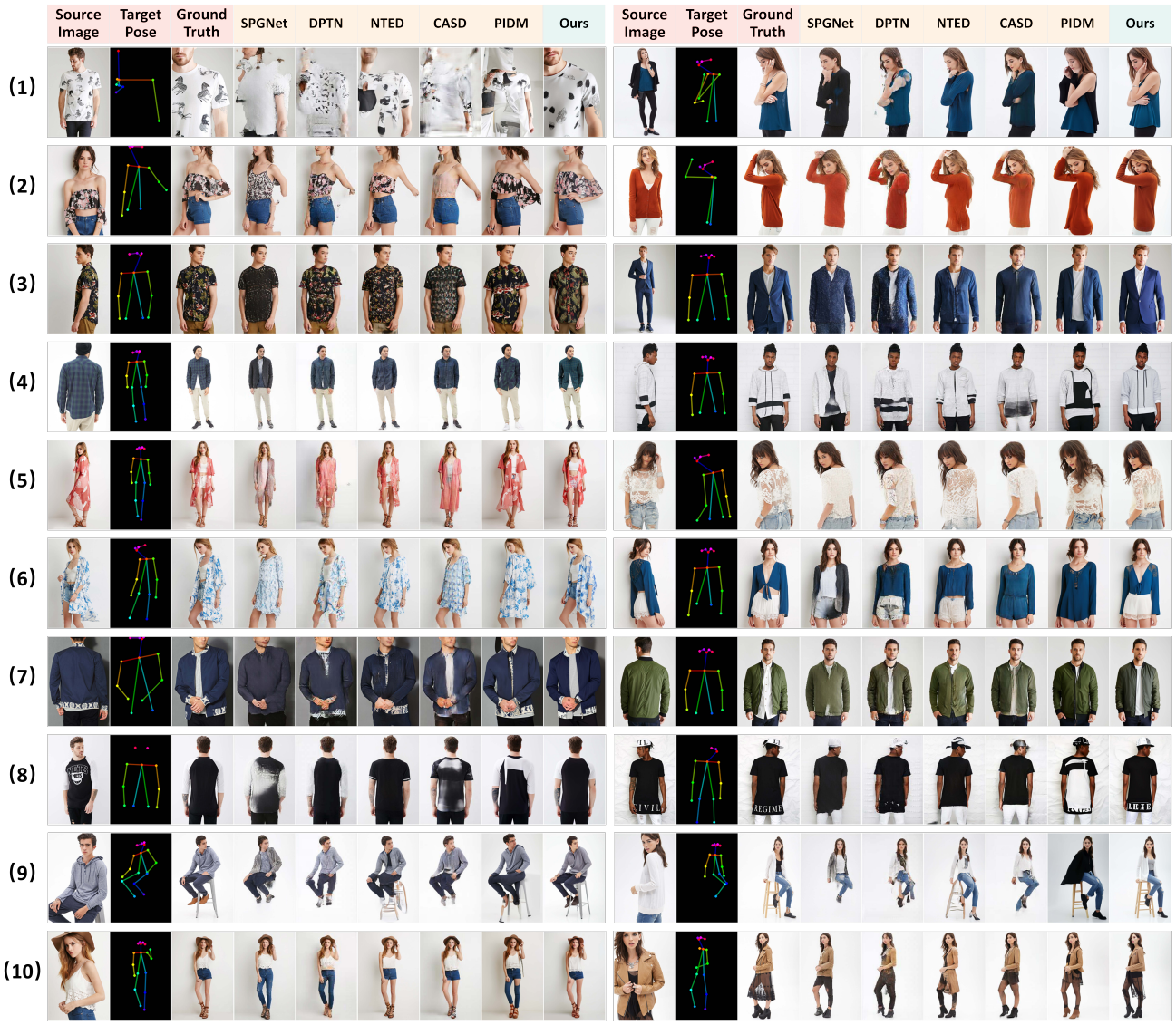

Figure 3. Qualitative comparisons with SPGNet [22], DPTN [48], NTED [31], CASD [53] and PIDM [1].

Implementation details. Our method is implemented with PyTorch [27] and HuggingFace Diffusers [43] on top of the Stable Diffusion [32] with the version of 1.5. The source image is resized to 256×256 and the source image encoder $\mathcal{H}_S$ is a standard Swin-B [21] pretrained on ImageNet [4]. The default settings and the number of trainable parameters in each component are summarized in Tab. 1. We train for 100 epochs using the Adam [16] optimizer with a base learning rate of 5e-7 scaled by the total batch size. The learning rate undergoes a linear warmup during the first 1,000 steps and is multiplied by 0.1 at 50 epochs. For classifier-free guidance, we set w_{pose} and w_{id} to 2.0 during sampling, and drop the condition $\boldsymbol{x}_s$ and $\boldsymbol{x}_p$ with a probability of $\eta = 20(\%)$ during training.

4.2. Quantitative Comparison

We quantitatively compare our method with both GAN-based and diffusion-based state-of-the-art approaches in terms of objective metrics. The evaluation is performed on both 256×176 and 512×352 resolutions as the same as in [9, 31]. As shown in Tab. 2, our method significantly outperforms the state-of-the-art across all metrics on both resolutions. In particular, compared to the other two diffusion-based methods [1, 9] in Tab. 1, we achieve better reconstruction with simpler 2D-only pose annotations and fewer trainable parameters. The metrics for VAE [7] reconstructions and the ground truths are also provided for reference. It is worth noting that the results we obtain by running with the checkpoint provided by PIDM [1] suffer from severe

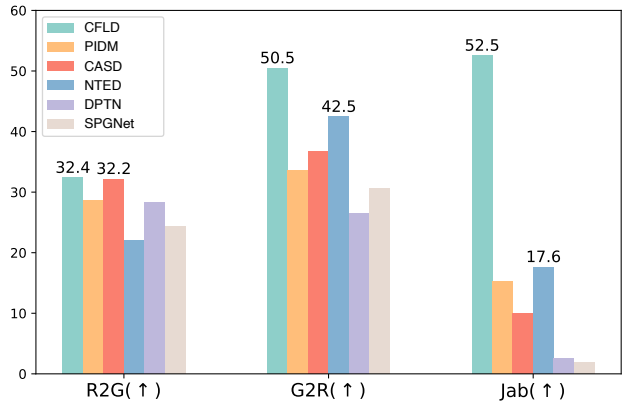

Figure 4. User study results in terms of R2G, G2R and Jab metrics.

overfitting, resulting in a large gap between the quantitative results of provided images and those from the checkpoint.

4.3. Qualitative Comparison

In Fig. 3, we present a comprehensive visual comparison with recent approaches that are publicly available and reproducible[1], including SPGNet [22], DPTN [48], NTED [31], CASD [53] and PIDM [1]. Our observations can be summarized as follows. **(1)** Both GAN-based and diffusion-based methods suffer from overfitting the human poses. When generating some target poses that are extreme or not common in the training set, existing methods show severe distortions as demonstrated in `rows 1-2`. Since we decouples the controls of fine-grained appearance and pose information, our method circumvents the potential overfitting problem and always generates a reasonable pose with the conditioning coarse-grained prompt and fine-grained appearance bias. **(2)** For source images in `rows 3-6` with more complex clothing, our generated images better preserve the textures details while aligning with the target pose thanks to the robust coarse-to-fine learning curriculum of hybrid-granularity attention module. For other methods, although they match in color, the clothes either exhibit blurring and distortion (SPGNet, DPTN, and CASD) or are spliced unnaturally in texture, creating a large gap from the source image (NTED and PIDM). **(3)** As for cases where the target pose requires visualization of areas invisible in the source image, our method exhibits strong understanding and generalization capabilities. With a semantic understanding of the source image provided by the perception-refined decoder, our method is aware of what should be predicted when the person turns around or sits down as illustrated in `rows 7-10`, such as different patterns on the front and back of clothes, the sitting chair, and lower body wear.

[1]The checkpoints and generated images of the related work PoCoLD [9] are not publicly available for qualitative comparison.

Method	Biasing	Trainable	Prompt	LPIPS↓	SSIM↑
B1		$\boldsymbol{K}, \boldsymbol{V}$	M-S	0.2018	0.6959
B2		$\boldsymbol{K}, \boldsymbol{V}$	CLIP	0.2099	0.6944
B3		$\boldsymbol{K}, \boldsymbol{V}$	PRD	0.1615	0.7293
B4		$\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$	PRD	0.1742	0.7198
B5	$\boldsymbol{Q}$	$\boldsymbol{K}, \boldsymbol{V}$	Swin	0.1912	0.7038
Ours	$\boldsymbol{Q}$	$\boldsymbol{K}, \boldsymbol{V}$	PRD	0.1519	0.7378

Table 3. Quantitative results for ablation studies. M-S is short for multi-scale fine-grained appearance features similar to [1, 9].

Figure 5. Qualitative ablation results. Our approach has a high-level understanding of the source image rather than forced alignment. It is also less prone to overfitting through the complementary coarse-grained prompts and fine-grained appearance biasing.

4.4. User Study

To verify the gap between generated and real images as well as our superiority over the state of the arts, we have recruited over 100 volunteers to perform the following two user studies following PIDM [1]. **(1)** For the R2G and G2R metrics, volunteers were asked to discriminate between 30 generated images and 30 real images from the testing set. Each volunteer could only see the generated images of a specific method, and the pairs of source image and target pose for generation were consistent across methods for a fair comparison. From the results in Fig. 4, chances of a real image being recognized as generated (R2G) are relatively low, and over half of the images we generated are recognized as real (G2R), demonstrating that our method generates more realistic images that are less likely to be judged as fake by humans. **(2)** For the Jab metric, each volunteer was asked to choose the best match to the ground truth from the generated images of different methods. Compared to other methods, our Jab score achieved 52.5 percent, significantly higher (+34.9) than the counterpart in second place, indicating that our method is more preferred and generates better texture details and pose alignment.

Figure 6. (a) Style transfer results of our method. The appearance in the reference image can be edited while maintaining the pose and appearance. This is achieved by masking out regions of interest in the reference image and requires no additional training. (b) The interpolation results show that texture details can be gradually shifted from one style to another in a smooth manner (from Style 1 to 2).

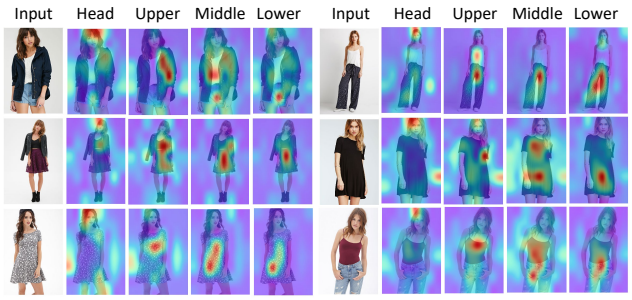

Figure 7. Visualizing attention maps by different queries of the prompt decoder. The maps are averaged over all attention heads.

4.5. Ablation Study

We perform ablation studies at multiple baselines to compare with our method. The quantitative results are presented in Tab. 3. `B1` is referenced from the other two diffusion-based approaches [1, 9] that incorporate multi-scale fine-grained appearance features as conditional prompts. we also experiment with CLIP image encoder [29] in `B2` to produce descriptive coarse-grained prompts for source images, which is first explored by an image-editing approach [45] that are also conditioned purely on images. Together with the qualitative results in Fig. 5, we can see that even very simple textures sometimes fail to be preserved, suggesting that these prompts are not compatible with the pre-trained SD model. To provide coarse-grained features that are more specific to person images, we integrate the Perception-Refined Decoder (PRD) into `B3`. The reconstruction metrics (i.e., LPIPS and SSIM) in Tab. 3 reveal a significant improvement in the quality of generated images, which validates the effectiveness of our proposed PRD. While this can be confirmed qualitatively in Fig. 5, there is still a lack of textural details as indicated by the red box. To address this issue, we experiment with training more parameters in the UNet as `B4` and instead observe a decrease in performance. This implies that the generalization ability of SD model is compromised, which is not our expectation. Thus we come up with the Hybrid-Granularity Attention (HGA) to bias the queries and achieve state-of-the-art results both quantitatively and qualitatively. In order to verify whether the source image encoder (i.e., Swin Transformer [21]) is able to learn sufficient information for HGA and give a useful prompt, we abandon the PRD in `B5`. The qualitative results in Fig. 5 demonstrate that both `B4` and `B5` are overfitting, only our method circumvents this problem by learning in a coarse to fine-grained manner.

Visualization. In Fig. 7, we visualize the effectiveness of different queries in $\mathcal{H}_D$. The attention maps reflect different human body parts of person images captured by learnable queries, which proves that we have a high-level understanding of the source images and thus less prone to overfitting.

4.6. Appearance Editing

Style Transfer. Our CFLD inherits the strong generation ability of SD model by freezing the vast majority of its parameters. Thus the style transfer can be achieved simply by masking without additional training. Specifically, we mark the regions of interest in the reference image $\boldsymbol{y}^{ref}$ as a binary mask $\boldsymbol{m}$. During sampling, the noise prediction is decomposed into $\epsilon'_t = \boldsymbol{m} \cdot \epsilon_t + (1 - \boldsymbol{m}) \cdot \boldsymbol{z}_t^{ref}$, where the ϵ_t is based on the pose from $\boldsymbol{y}^{ref}$ and the appearance from different styles of source images. Let $\boldsymbol{z}_t^{ref}$ be the noisy latent at timestep t mapped from $\boldsymbol{z}_0^{ref} = \mathcal{E}(\boldsymbol{y}^{ref})$ as in Eq. (1). From the results in Fig. 6(a), our method generates realistic and coherent texture details in the regions of interest.

Style Interpolation. Additionally, our CFLD supports arbitrary linear interpolation of both coarse-grained prompts and fine-grained appearance biases. As shown in Fig. 6(b), our generated images are faithfully reproducing different styles with smooth transitions.

5. Conclusion

This paper presents a novel Coarse-to-Fine Latent Diffusion (CFLD) method for Pose-Guided Person Image Synthesis (PGPIS). We circumvent the potential overfitting problem by decoupling the fine-grained appearance and pose information controls. Our proposed Perception-Refined Decoder (PRD) and Hybrid-Granularity Attention module (HGA) enable a high-level semantic understanding of person images, while also preserving texture details through a coarse-to-fine learning curriculum. Extensive experiments demonstrate that CFLD outperforms the state of the arts in PGPIS both quantitatively and qualitatively. Our future work will investigate whether the CFLD can be extended to more general image synthesis tasks besides PGPIS.

References

[1] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *CVPR*, page 5968–5976, 2023. 2, 4, 5, 6, 7, 8

[2] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023. 5

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, page 7291–7299, 2017. 1, 5

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, page 248–255, 2009. 6

[5] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 3

[6] Patrick Esser and Ekaterina Sutter. A variational u-net for conditional appearance and shape generation. In *CVPR*, page 8857–8866, 2018. 1, 2

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, page 12873–12883, 2021. 2, 3, 4, 6

[8] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, pages 7297–7306, 2018. 2, 5

[9] Xiao Han, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, and Tao Xiang. Controllable person image synthesis with pose-constrained latent diffusion. In *ICCV*, page 22768–22777, 2023. 2, 4, 5, 6, 7, 8

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, page 770–778, 2016. 4

[11] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, page 16000–16009, 2022. 2

[12] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv:2208.01626*, 2022. 3

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 5

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshops*, 2021. 3, 5

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[17] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *CVPR*, page 3693–3702, 2019. 1, 2

[18] Vivian Liu and Lydia B Chilton. Design guidelines for prompt engineering text-to-image generative models. In *CHI*, pages 1–23, 2022. 1

[19] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *ICCV*, pages 5904–5913, 2019. 1, 2

[20] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, page 1096–1104, 2016. 2, 5

[21] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, page 10012–10022, 2021. 4, 5, 6, 8

[22] Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, and Wangmeng Zuo. Learning semantic person image generation by region-adaptive normalization. In *CVPR*, page 10806–10815, 2021. 1, 2, 5, 6, 7

[23] Liqian Ma, Xu Jia, Qianru Sun, B. Schiele, T. Tuytelaars, and L. Gool. Pose guided person image generation. In *NeurIPS*, 2017. 2, 5

[24] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *CVPR*, pages 99–108, 2018. 2

[25] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *CVPR*, page 5084–5093, 2020. 1, 2, 5

[26] Chong Mou, Xintao Wang, Liangbin Xie, Jing Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv:2302.08453*, 2023. 2, 3, 4, 5

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32, 2019. 6

[28] Nikita Pavlichenko and Dmitry Ustalov. Best prompts for text-to-image models and how to find them. In *SIGIR*, pages 2067–2071, 2023. 1

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, page 8748–8763, 2021. 3, 8

[30] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H. Li, and Ge Li. Deep image spatial transformation for person image generation. In *CVPR*, page 7690–7699, 2020. 1, 2, 5

[31] Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H. Li. Neural texture extraction and distribution for controllable person image synthesis. In *CVPR*, page 13535–13544, 2022. 1, 2, 5, 6, 7

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, page 10684–10695, 2022. 1, 3, 6

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, page 234–241, 2015. 3

[34] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *NeurIPS*, 29, 2016. 5

[35] Kripasindhu Sarkar, Vladislav Golyanik, Lingjie Liu, and Christian Theobalt. Style and pose control for image synthesis of humans from a single monocular view. *arXiv:2102.11263*, 2021. 1, 2

[36] Aliaksandr Siarohin, Enver Sangineto, Stephane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *CVPR*, page 3408–3416, 2018. 1, 5

[37] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 3, 5

[38] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 3

[39] Hao Tang, Song Bai, Li Zhang, Philip H. S. Torr, and Nicu Sebe. Xinggan for person image generation. In *ECCV*, page 717–734, 2020. 1, 2

[40] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, pages 1921–1930, 2023. 3

[41] Leonid Nisonovich Vaserstein. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969. 5

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[43] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 6

[44] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *TIP*, 13(4):600–612, 2004. 5

[45] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *CVPR*, page 18381–18391, 2023. 8

[46] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. In *CVPR*, page 7982–7990, 2021. 1, 2, 5

[47] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, page 3836–3847, 2023. 1, 2, 3, 4

[48] Pengze Zhang, Lingxiao Yang, Jianhuang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *CVPR*, page 7713–7722, 2022. 1, 4, 5, 6, 7

[49] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, page 586–595, 2018. 5

[50] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, page 16816–16825, 2022. 1

[51] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 1

[52] Xingran Zhou, Bo Zhang, Ting Zhang, Pan Zhang, Jianmin Bao, Dong Chen, Zhongfei Zhang, and Fang Wen. Cocosnet v2: Full-resolution correspondence learning for image translation. In *CVPR*, page 11465–11475, 2021. 5

[53] Xinyue Zhou, M. Yin, Xinyuan Chen, Li Sun, Changxin Gao, and Qingli Li. Cross attention based style distribution for controllable person image synthesis. In *ECCV*, page 161–178, 2022. 1, 2, 5, 6, 7

[54] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *CVPR*, page 2347–2356, 2019. 1, 2, 5